DOCUMENT RESUME

ED 169 099 TM 008 496

AUTHOR Childs, Roy
TITLE Norm Referenced Testing and the Standard Scores. Basic Testing Services.
INSTITUTION National Foundation for Educational Research Publishing Co., London.
PUB DATE [76]
NOTE 13p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Calculation; Foreign Countries; *Mathematical Formulas; Mathematical Models; Measurement; Norm Referenced Tests; *Norms; *Scores; Scoring Formulas; Standardized Tests; Statistical Analysis
IDENTIFIERS Great Britain; National Foundation for Educational Research; *Normal Curve

ABSTRACT

The norm-referenced score scale used by the National Foundation for Educational Research (NFER) is described. The usefulness of standardized scores is explained by a simple numerical example, and the formulas and computations are shown for calculating a mean, a standard deviation, and a deviation or z score. The need for a representative sample is explained and the normal distribution curve is described. Finally, NFER standardized scores and percentiles are explained, tabulated and presented graphically. (CTM)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

- ED169099

# Basic Testing Services
# NORM REFERENCED TESTING
# and the standard scores

by Roy Childs (Psychologist)
N F E R Publishing Company

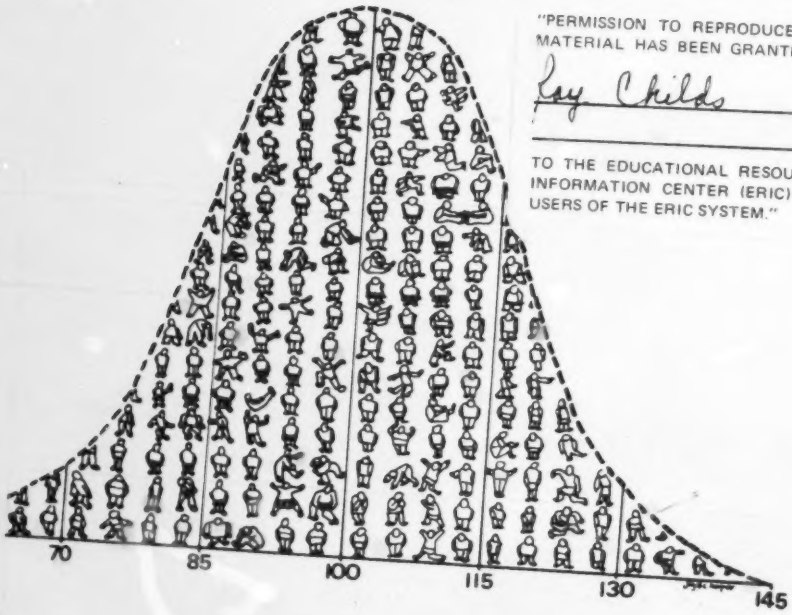
U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Roy Childs

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."




TM008 496

## Introduction

Teachers are manifestly capable of producing their own tests. They can choose those items* which are appropriate to the particular syllabus they have devised and within the capability of their pupils.

Why then, are standardized tests being used so widely in schools? The answer lies in the much wider interpretation that can be made from a standardized test and this, in turn, relies very much on the nature of the standard score. It is hoped that the following paragraphs will enable a teacher to understand the nature of standard scores and therefore to extract the maximum information from such a test while still being aware of the limitations of such scores for making a more general assessment of a child or group of children.

## Teacher-Made Tests

Classroom tests made by the teacher can be very useful for certain purposes. They can be used for ranking children in order from highest to lowest in terms of achievement. They can also help to evaluate the success of a particular course. However, the scores obtained on such tests have certain specific limitations. They do not enable us to:

1) compare performance in different subjects or even different topics in the same subject. Therefore they do not enable us to obtain an overall picture which would highlight discrepancies in achievement in different subjects.
2) combine scores on different tests since, for example an easy test with 100 items would swamp the results of a difficult test with 10 items.
3) compare performance in relation to some kind of standard or norm.
4) make an estimate as to whether differences between scores are really significant or whether they are due to an inherent unreliability in the test itself.
5) monitor progress over time because of the difficulties mentioned above.

---

* The term "item" is used for each individual question or unit which can be marked pass or fail. A test is therefore made up of a number of items.

These limitations may become more evident if we consider a simple analogy. We could measure the height of the pupils in our class by using a bundle of rods all of different lengths. We can then take each rod and decide whether a pupil is taller or shorter than the rod. Obviously the *taller* pupils will be taller than a greater number of rods as compared with the shorter pupils. This would be the same as the more *able* pupils passing more of the items in any test than the less able pupils. We can therefore obtain a measure of height by counting the number of rods that the pupil "passes" in the same way as a measure of ability can be obtained by counting the number of items that a pupil passes or gets right. This analogy may make the limitations of this kind of assessment clearer. For example the pupils can be ranked from tallest to shortest by using the rods, but the accuracy with which this can be done will depend on how finely graded the rods are i.e. whether each rod is only a few millimetres longer or several centimetres longer than the previous one.

Furthermore the limitations 1, 2 and 5 all involve administering at least one other test. In our analogy the equivalent to this would be to measure height using another batch of rods of different lengths. However, we would now obtain a different score since the score is dependent on the number of rods at each particular length. Although we would still expect a similar rank order, it can be seen that the scores obtained from simply adding the number of rods passed (i.e. the raw scores) on each batch are not directly comparable. A similar re-test problem is apparent when attempting to monitor progress over time. The children will soon "outgrow" the rods and a new batch will therefore be needed. But scores on this new batch cannot be related to the previous scores.

**Test Scores**

It can be seen that the use of raw scores (i.e. the number of items right) imposes severe limitations on the way we can interpret test results. This can be demonstrated by looking at the test results of the following group of children on a battery of 3 tests (see Table 1).

What can we say about the performance of any of these children? Obviously the score on any of these tests depends not only on the competence or ability of the child, but also on the difficulty of the items and the total number of items in the test. Is John's performance of 60 in Maths better than his score of 51 in Verbal Reasoning or of 45 in English? It is certainly his highest score, but at the same time it is his lowest rank. A teacher can look at each score

**Table 1**

| | *Maths Score* | *(rank)* | *English Score* | *(rank)* | *Verbal Reasoning Score* | *(rank)* |
|---|---|---|---|---|---|---|
| John | 60 | (4) | 45 | (2) | 51 | (3) |
| Charles | 56 | (5) | 35 | (3) | 83 | (1) |
| Egbert | 64 | (3) | 50 | (1) | 36 | (5) |
| Anne | 68 | (2) | 32 | (4) | 80 | (2) |
| Ruth | 70 | (1) | 28 | (5) | 30 | (6) |
| Claudia | 42 | (6) | 20 | (6) | 50 | (4) |

and compare it to every other score in a series of operations, but this procedure is tedious and, fortunately, unnecessary. There are methods of comparing John's score to some meaningful reference points whereby the teacher can interpret the score directly. These methods form the basis of the Standard Scores.

**Making Scores Comparable**

To make scores directly interpretable and comparable we need to relate them to some meaningful reference points and one such reference is the average or the mean of a particular group. This, at least, enables us to judge a person's score in relation to the average. In this example the mean score in Maths is:

$$(60 + 56 + 64 + 68 + 70 + 42) \div 6 = 360 \div 6 = 60$$

John's mathematics score of 60 is therefore average for this group. Similarly the mean score in English is 35 and in Verbal Reasoning is 55. This tells us that John's English score of 45 is considerably higher than the average whereas his Verbal Reasoning score of 51 is slightly below average. However, this way of describing test scores is not sufficient since we do not know what "considerably better" or "slightly worse" really mean. We need to develop a way to quantify these terms.

A closer look at the test scores reveals that the Verbal Reasoning scores have a much wider scatter about the mean than either English or Maths. In fact the scores range from 30 to 83 (i.e. 50 points) whereas for English and Maths the range is only 30 points and 28 points respectively. It would therefore seem reasonable to suggest that a person would have to score considerably more points above the mean in Verbal Reasoning to be doing as well as someone with

only a few points above the mean in Maths. This may seem a debatable assumption since the results may indicate a genuine difference in the spread of reasoning ability compared to the spread of Maths attainment. However, the larger the groups we deal with, the more confidence we would have in this assumption. Since most standardized tests use large, representative samples, the assumption would appear justifiable.

We therefore need a measure of *spread*. The most commonly used measure is the Standard Deviation (S.D.) since it is very stable and not excessively influenced by only one or two extreme scores. It can be calculated as follows:

**Table 2**

| | *I* *Matbs* | *II* *Deviation from the mean of 60* | *III* *Deviation Squared* |
|---|---|---|---|
| John | 60 | 0 | 0 |
| Charles | 56 | −4 | 16 |
| Egbert | 64 | +4 | 16 |
| Anne | 68 | +8 | 64 |
| Ruth | 70 | +10 | 100 |
| Claudia | 42 | −18 | 324 |
| *Total* | 360 | 0 | 520 |

The first step is to subtract each person's score from the mean. This tells us "how far" his score is from the mean. The sum of these deviations is always zero (see column II) which is not a very useful measure of spread! In order to "remove" the minus signs, each deviation is squared as in column III. The average of these squared deviations is then found to be:

$$520 \div 6 = 86.67$$

However, this figure is on a different scale from the original test scores since we squared all the deviations. To return to the original scale we find the square root and it is this which finally yields the Standard Deviation (S.D.).

$$\text{S.D.} = \sqrt{86.67} = 9.3^{*}$$

* The formula is therefore S.D. $= \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$ where $x$ is the score, $\bar{x}$ is the mean and $n$ is the number in the group. The reader will often find $n - 1$ is used rather than $n$ since this provides an *unbiased* estimate of spread. Reference (2) deals with the subject in detail.

A similar calculation for English gives an S.D. of 10.1 and for Verbal Reasoning an S.D. of 20.1. We can now quantify scores which differ from the mean in terms of the S.D. We can also assume that a person who scores 1 S.D. above the mean in Maths is performing at the same level as someone who is performing 1 S.D. above the mean in Verbal Reasoning. In other words, a score of about 69 (mean of 60 + S.D. of 9.3) in Maths is roughly equivalent to a score of 75 (mean of 55 + S.D. 1) in Verbal Reasoning. We can express this another way. If we take a score and subtract the mean and divide by the S.D. we get the score expressed as the number of S.D.s from the mean:

$$\frac{\text{Score} - \text{Mean}}{\text{S.D.}}$$

Thus the score of 69 in Maths becomes: $\frac{69 - 60}{9.3} = 0.97$ S.D.s

A score of 46 in Maths becomes: $\frac{46 - 60}{9.3} = -1.5$ S.D.s

**Obtaining the Standard Scores**

We now have a scale (called the z-scale) which is meaningful and consistent for most tests we care to use. However, the problem with the z-scale is that it has a mean of zero, (c.f. $\frac{\text{mean} - \text{mean}}{\text{S.D.}} = 0$) and has scores which have both minus signs and decimal points. Since this is not the clearest way to present scores, there are various conventions used to convert the z-scale to another more easily understood scale. They are all based on the fact that any constant can be added to the mean without altering the shape of the distribution. In other words the S.D. will remain the same even if a constant number is added to each score. However, if all the scores were multiplied by a certain constant, the S.D. would increase by the same factor (i.e. S.D. x constant). We can therefore transform a distribution of scores to a distribution having a particular mean and S.D.

The NFER usually transforms its distributions so that the mean is equal to 100 and the S.D. is equal to 15. In other words, someone who scores exactly the mean score will be allocated a score of 100. A

person who scores exactly 1 S.D. above the mean will be allocated a score of 115. Intermediate positions are calculated according to a strict mathematical formula (Lawley, 1950).

**Interpretation of Standard Scores**

The interpretation of standard scores is based on a comparison with a particular group. Our interpretation is therefore completely tied to that group. This must always be borne in mind since strange interpretations could otherwise result. To return to our stick analogy, if the mean and S.D. were calculated for a group of pygmies, then the score for a British person would place them way above the mean. However, this would not tell us whether the person was tall or short compared with the British population. A test must therefore be standardized (the term used for selecting a group on which the standard scores are based) on a group that is relevant to the interpretations required. Most standardized tests are given to a sample which is intended to be representative of the national population. Generally speaking, therefore, the larger the sample, the more confident we would be that it is truly representative. However, sampling is not always as accurate or as rigorous as we would want and so, in some cases, it may be true that a smaller sample reflects the national picture more accurately.

For certain purposes it is not particularly useful to make a comparison with a national sample. Inner city schools, for example, may find it more appropriate to use a test which has standard scores obtained on a similar inner city area. These local norms can often be obtained from the test publishers if requested.

**Further Interpretation**

If we again consider measuring height we can predict that we would find more people around the average height than at the extremes of tallness and shortness. In fact, for a small group of about 40 people taken at random we would expect a distribution of people with different heights to look like Figure 1.

If we were to measure many more people even more accurately, we would get a distribution which would smooth out to look like Figure 2.

This kind of distribution is very commonly found in measurement and is called the normal distribution. In testing we make the

*Figure 1*

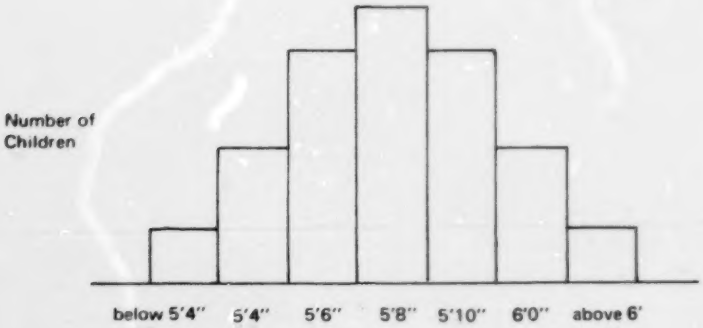
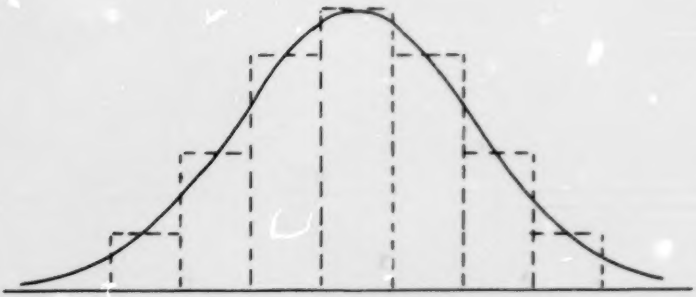



*Figure 2*

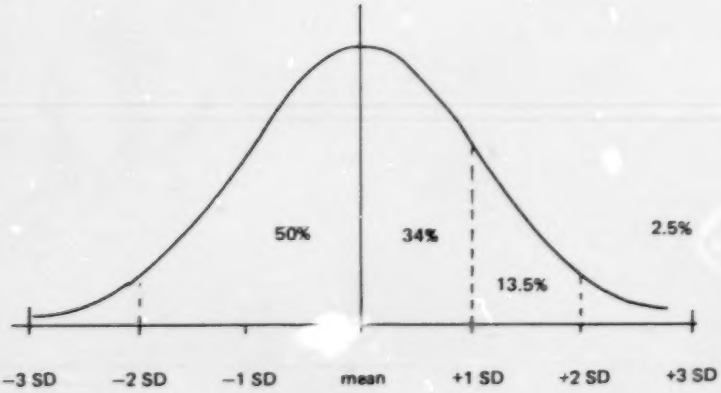


assumption that ability or achievement will also be normally distributed. There is no way of proving that this is so for two reasons:

(a) We cannot directly observe things such as ability or achievement.
(b) The distribution of scores would be altered if a different set of test items were selected.

However, a normal distribution would appear to be a reasonable assumption and has important implications. It can be defined very precisely in mathematical terms and there is a very precise relationship between the distance from the mean and the proportion of people who would obtain more or less than that particular score. This relationship can be demonstrated as follows.

*Figure 3*




Someone who scores at 1 S.D. above the mean will be scoring better than 84% (50% + 34%) of the population. In fact, any number or fraction of S.D.s above or below the mean corresponds to a precise percentile position. It is this relationship which can be used to allocate a standard score for each raw score obtained. As mentioned earlier, the NFER tests use standard scores based on a mean of 100 and an S.D. of 15*.

* The NFER standard scores are not, in fact calculated in this way. Since an allowance is made for age, it would take enormous samples to obtain a smooth normal distribution. There are often fluctuations in the numbers of children in each age band. A complex analytical technique is therefore used to smooth the curve and produce standardized scores. The interested reader is referred to (5) in the list of references.

The percentile position for each standard score can be read off from the table below.

**Table for Converting**
**NFER Standardized Scores to Percentiles**

| *Standardized Score* | *Percentile* | *Standardized Score* | *Percentile* | *Standardized Score* | *Percentile* |
|---|---|---|---|---|---|
| 140 | 99.7 | 115 | 84.9 | 91 | 28.5 |
| 139 | 99.6 | 114 | 83.3 | 90 | 26.3 |
| 138 | 99.5 | 113 | 81.6 | 89 | 24.2 |
| 137 | 99.4 | 112 | 79.7 | 88 | 22.2 |
| 136 | 99.3 | 111 | 77.8 | 87 | 20.3 |
| 135 | 99.1 | 110 | 75.0 | 86 | 18.4 |
| 134 | 98.9 | 109 | 73.7 | 85 | 16.7 |
| 133 | 98.7 | 108 | 71.5 | 84 | 15.1 |
| 132 | 98.5 | 107 | 69.1 | 83 | 13.6 |
| 131 | 98.2 | 106 | 66.7 | 82 | 12.2 |
| 130 | 97.9 | 105 | 64.3 | 81 | 10.9 |
| 129 | 97.5 | 104 | 61.8 | 80 | 9.7 |
| 128 | 97.1 | 103 | 59.2 | 79 | 8.6 |
| 127 | 96 7 | 102 | 56.6 | 78 | 7.6 |
| 126 | 96.1 | 101 | 54.0 | 77 | 6.7 |
| 125 | 95.5 | 100 | 51.3 | 76 | 5.9 |
| 124 | 94.9 | 99 | 48.7 | 75 | 5.1 |
| 123 | 94.1 | 98 | 46.0 | 74 | 4.5 |
| 122 | 93.3 | 97 | 43.4 | 73 | 3.9 |
| 121 | 92.4 | 96 | 40.8 | 72 | 3.3 |
| 120 | 91.4 | 95 | 38.2 | 71 | 2.9 |
| 119 | 90.3 | 94 | 35.7 | 70 | 2.5 |
| 118 | 89.1 | 93 | 33.3 | | |
| 117 | 87.8 | 92 | 30.9 | | |
| 116 | 86.4 | | | | |

The reader may expect a score of 100 to cut the distribution exactly in half and thus give a percentile position of 50. However, since a substantial number of people will score 100 it is only the middle person of the group who will be at the 50th percentile. We make the assumption that a person who scores 100 will be at the top end of all those with the same score. He is therefore placed at the 51.3 percentile position rather than the 50th.

## Conclusion

It is hoped that the reader will now appreciate the meaning of the standard scores obtained from our standardized tests. Not all standard scores are expressed using a mean of 100 and an S.D. of 15. However, most of them are based on a transformation of the z-scale to a more easily understood scale. Some of the most common standard scores are shown overleaf in relation to the normal distribution.

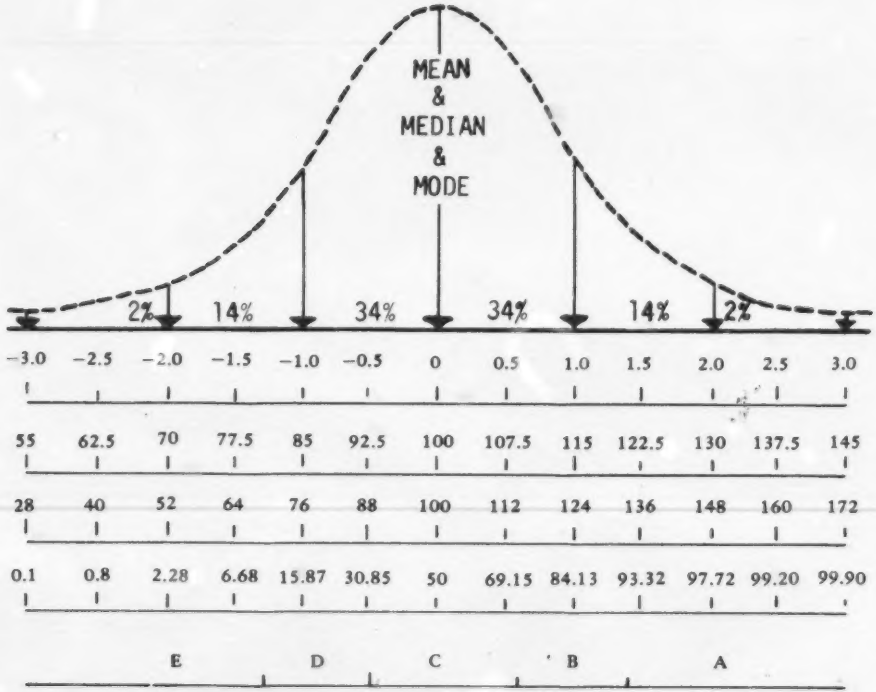
It is worth bearing in mind that Mental Ages (M.A.) and Reading Ages (R.A.) are not based on the z-scale. In fact the S.D. is not usually quoted since, at least at the later age levels, the distribution of M.A.s and R.A.s are not normally distributed and the S.D. varies for each age group tested. One year's retardation at one age level has a different significance from one year's retardation at another age level. This ambiguity in the meaning of M.A.s and R.A.s has led most test constructors to abandon this way of presenting scores, although it is still quite popular for reading tests.

Standard scores still have their limitations, however. In physical measurement, for example we would be unhappy with only being able to measure height by using a standard score based on a sample of other people. Standard scores, in that sense, are not true objective measurements since they are what we call "sample bound". In an attempt to obtain scores which are not sample bound certain statistical developments have taken place which are being considered in relation to Item Banking. But this is the subject of another document. It is certain, however, that standard scores, as explained here, will be with us for some time yet.

## References

(1) Crocker, B. *Statistics for the Teacher*. NFER, 1974.
(2) Freund, J.E. *Mathematical Statistics*. Prentice-Hall, 1963.
(3) Hoel, P.G. *Introduction to Mathematical Statistics*. Wiley International, 1946, 1971.
(4) Hudson, B. *Assessment Techniques*. Methuen.
(5) Lawley, D.N. "A Method of Standardizing Group Tests", *British Journal of Psychology* (Statistical Section), Vol. III, part II, June, 1950.
(6) Nunnally *Psychometric Theory*. McGraw-Hill, 1967.
(7) Thorndike and Hagen *Measurement and Evaluation in Psychology and Education*. Wiley International, 1955, 1961, 1969.
(8) Tuckman, B.W. *Measuring Educational Outcome*. Harcourt Brace Jovanovich Inc, 1975.
(9) Vincent, D. and Cresswell, M. *Reading Tests in the Classroom*. NFER, 1976.

Figure 4


MEAN
&
MEDIAN
&
MODE
2%
14%
34%
34%
14%
2%
−3.0 −2.5 −2.0 −1.5 −1.0 −0.5 0 0.5 1.0 1.5 2.0 2.5 3.0
Z-Scale
55 62.5 70 77.5 85 92.5 100 107.5 115 122.5 130 137.5 145
NFER Standard Scores
(S.D. =15)
28 40 52 64 76 88 100 112 124 136 148 160 172
I.Q. on Culture Fair Tests
(S.D. =24)
0.1 0.8 2.28 6.68 15.87 30.85 50 69.15 84.13 93.32 97.72 99.20 99.90
Percentiles
E D C B A
Grades